福井大学
工学部 研究報告
第41巻 第2号
1993年9月

英文コーパス検索における語の文法機能の利用

臼井 秀宣* 高濱 徹行** 小高 知宏**
舘 清隆*** 小倉 久和**

Retrieval System for English Corpus Based on the Grammatical Function of Word

Hidenobu USUI, Tetsuyuki TAKAHAMA, Tomohiro ODAKA,
Kiyotaka TACHI, and Hisakazu OGURA

(Received Aug. 31, 1993)

We execute retrieval of words or phrases for a variety of purposes making use of retrieval systems. A word in a sentence, generally, has the function(s) which can be defined in relation to other elements of the sentence. In this paper we will refer to this sort of function as grammatical function of word. And sometimes it is the case that a word can fulfill different functions in different sentences. Thus the word "there" is used either as a deictic locative adverb or as an expletive element. The problem for a retrieval system is the following: when we are in search of expletive "there" sentences, for example, the sentences containing locative "there" are irrelevant. It is necessary to decrease the number of these unexpected sentences in the result of retrieval, possibly up to zero. We have attempted to incorporate this feature into the retrieval system for the database called English Corpus by making use of grammatical function of word. This paper is a preliminary discussion of a prototype retrieval system applied to the English textbooks used at Japanese junior high schools.

1 はじめに

近年、自然言語で表された文章を対象とした全文データベースが広く使われるようになってきた。科学的文献や新聞記事などのテキスト型データベースでは、検索は、検索語句の記号列、文字列としての一致によっている。もちろん、検索文字列の完全一致ばかりではなく、前方一致や後方一致、包含などという部分一致や、同義語辞書や類語辞書によるある種の連想機構的な検索機能など、高速な一致検索とともに、高度な技術も利用されている。

* 大学院情報工学専攻
** 工学部情報工学科
*** 教育学部英語科

ところで、英語の研究・教育において、標準となる例文や用例を作成したり、表現の妥当性をチェックしたりすることは、重要なことである。しかし、英語を母国語としない研究者や教育者にとって、このことは簡単ではない。このような目的で容易に利用できるものに、英文コーパスがある。現代英語の全体像を把握し、研究・教育の対象とするべく、大規模な現代英文コーパスがいくつか編纂され、作成・提供されている。このような英文コーパスを利用した研究が少しずつ進んできている。

いくつかの英文コーパスがいままでに作成されているが、例として、アメリカ英語のブラウンコーパスについて簡単に述べておく。ブラウンコーパスは、ブラウン大学のネルソン・フランシスによって1963年から1964年の間に編集された。このコーパスは現代アメリカ英語の全体的姿を反映するものとなるように学術論文や科学小説など多くのジャンルの出版物から資料が採集された。資料の採取先は1961年にアメリカで出版されたものに限られている。様々な出版物から約2000語の連続を1テキストとし、500テキストが採取された。約百万語から成る、6.8メガバイトのテキスト型ファイルである。このコーパスは、1961年にアメリカで出版されたものをもとにして編集されているが、書き言葉は、急速に変容する話し言葉に比べて、比較的安定しており、今日の書き言葉における日常英語の資料とするのにそれほど抵抗はないものと考えられている。

英文コーパスのような英文データベースを、英文そのものの研究あるいは教育に利用するには、現在のような文字列の一致のみを利用した検索システムでは不十分である。文中の語句は他の語句から独立している訳ではなく、逆に密接に関連しているのであり、検索者の対象とする語句も、実は、検索者の意図としては明示してはいないが何らかの他の語句との関連、統語上の機能をもったものである。そして、検索者はその関連とともに対象の語句を検索しているのである。従って、文字列一致の検索によって得られたもののなかには、意図した関係をもっていない、対象外の用例、いわゆる検索ノイズが多数含まれることになる。その結果、検索者は、多くの検索結果を含んだ出力リストの中から意図したものを再び探すことになるのである。

語句がもつ他の語句との関係のうち、統語構造に関する関係、各語句の文法的な役割や係り結びなどの関係は、比較的検討しやすいと思われる。このような関係を、われわれは、語あるいは句の文法機能と呼ぶが、検索時に、検索語と同時にこの文法機能についても検索対象とすることで、検索者の意図しない用例を減少させ、検索のヒット率を向上させることができる。われわれは、このような意図に基づき、英文コーパスに対する検索システムの検討を行っている。現在、このような検索機能の可能性を検討するため、本格的な英文コーパスではなく、構文的に比較的単純であると思われる中学校の英語教科書の英文を対象として、プロトタイプシステムを作成してきている。本論文では、このような試みの結果について報告する。

2　語の文法機能を利用した英文コーパス検索システム

計画している検索システムの構成を説明する。図1はシステム構成の概要図である。

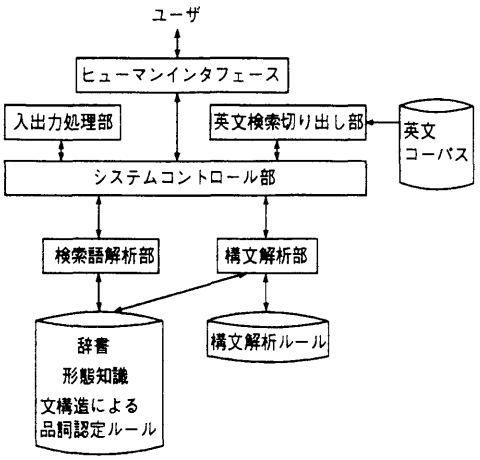

図1:システム構成の概要図

1. 入出力処理部において、検索する語句とその文法機能の指定を入力として受けとり、入力のチェックを行った後、検索語解析部へ送る。
2. 検索語解析部では、辞書、形態知識、文構造による品詞認定ルールを利用して検索語句の文法機能（たとえば品詞情報など）を得て、英文検索切出し部へ送る。
3. 英文検索切出し部では、語句によりコーパスの検索を行い、検索語句を含む文を切出し、構文解析部に送る。
4. 構文解析部では、切出された文の語句・形態素解析と構文解析を行い、検索語句の機能情報とのマッチングをチェックする。マッチすれば結果を入出力処理部へ送り、引き続き英文検索切出し処理を続ける。
5. 最後に、入出力処理部では、得られた語句を含む、切出された英文を出力する。

システムコントロール部は、以上の各部の辞書や知識ベースを管理するとともに、ヒューマンインタフェースを受け持つ。

英文コーパスを対象としたこのようなシステムを研究目的で使うためには、システムの適切な応答速度、対象をのがさず（第一種の過誤を少なく）ノイズを少なく（第二種の過誤を少なく）することが重要である。また、このようなシステムによる教育支援を考えると、英語の文章作成あるいは論文作成において、語句の具体的使用法のチェックなど広範な利用がある。この時には、使いやすいユーザインタフェースの充実が不可欠となろう。

3 プロトタイプシステムの作成

3.1 プロトタイプシステムの構成

言語学上の研究に利用される英文コーパスを対象として、語の文法機能による検索システムの構築が課題である。われわれは、そのような課題の前課題として、まず、初級英文、中学英語で利用される

教科書の英文を対象に、語の文法機能を用いたプロトタイプ検索システムを作成し、基礎的な研究を行なっている。中学英語の教科書は、コーパスのようなデータベースと比べると、小規模であり、また複雑な構造の文がほとんどない。そのため、辞書の作成やシステムによる構文の解析が容易になると思われる。用いた教科書は、東京書籍のニューホライズンである。原則として原文をそのまま用いたが、Mr. や Mrs. などの単語に付く文字"." は、文の終りを示す文字（ピリオド）"." と容易に区別するために、"&" に置き換えている。

プロトタイプシステムの構成は、前の節で説明したシステムの構成とほぼ同じであるが、いくつかの機能ははぶいてある。処理部の機能の変更点と実現されていない機能を以下に示す。

1. 入出力処理部において、検索する語や句とその文法機能を入力として受けとるが、そのチェックを行なっていない。プロトタイプシステムを作成している現段階では、正しい入力がなされることを前提としている。
2. 検索語解析部は辞書、形態知識、文構造による品詞認定ルールを利用してユーザが入力した検索語・句の機能を確定する処理部であるが、現在は、それらを利用していない。入力は正しくされていることを仮定しているためである。
3. 英文検索切り出し部は、検索対象ファイルの英文データベースから検索語を含む英文を一文切り出す毎に、構文解析部へその英文を送ることになっている。しかし、プロトタイプシステムでは、システムの構成を簡単にするため、検索語を含むすべての英文を検索対象ファイルの英文データベースから切り出した後、構文解析部でそれらの英文を解析する。

英文データベースからの英文検索切り出し部の検索は、高速性を要求されるため、検索はC言語を用いてBM (Boyer-Moore) 法を使った。英文を切り出すときに、切り出す英文に kcl(kyoto common lisp) システム上で定義済みの記号 ' や , などが含まれる場合、それらを別な記号にそれぞれ置き換えている（表1）。

定義済みの記号	置き換え記号
“	%
’	—
,	@

表1．定義済み記号の置き換え

形態素解析や構文解析をはじめ、全体のシステムは、現在のところ kcl で作成しており、英文検索切出し部では、kcl とCのプログラムをリンクしている。構文解析は LL(1) 法に依った。用いた構文規則は 3.3 節で説明するが、全体の規則は付録にまとめておいた。

3.2 辞書構成

辞書は単語の品詞を決定する形態素解析で用いられる。辞書の構成は、見出し語とその語が有する品詞をリスト形式にしたものである。見出し語とその品詞は中学生の英語教科書で使われている語のも

のである。語の形態の特徴により品詞を認定できる場合は、その語を辞書に登録していない。これらの単語の品詞の認定については次章で述べる。辞書では、複数の品詞を有する語に対処するため、品詞自身もリスト形式としてある。

次は辞書の一部分である。

```
(Africa (@n))
(African (@adj))
(after (@prep @conj))
(afternoon (@n))
```

@の付いた文字列は品詞を表す。

@n :名詞 @adj :形容詞
@prep:前置詞 @conj:接続詞

本研究の辞書に含まれている品詞の種類は全部で16種類である。その全品詞は付録に示した。

構文解析を容易にするため、英和辞書などに掲載されている品詞とは異なる品詞を割り当てた単語や、特別の品詞をあてたものがいくつかある。例えば、人称代名詞の所有格の my, your の単語の品詞は代名詞所有格であるが、形容詞とした。あるいは、He's, I've などは主語＋動詞、主語＋助動詞という品詞とした。このような品詞は必要に応じて適宜導入する。

単語の mustn't, can't, don't, doesn't, isn't, aren't は、語尾に 't が付いており、形態において特徴のある単語となっているが、これらの単語もそれぞれ辞書に登録した。

3.3 構文解析ルール

構文解析はLL(1)法に依った。構文解析ルールは、ルール番号、非終端記号、導出される記号（群）、先読み集合のリストからなるリスト形式で表現している。本研究で用いているルールのいくつかを以下に示す。

```
(1 s sub vp (@n @def @indef @pron @adj))
(2 s sub vp com (@n @def @indef @pron @adj))
(3 s sub vp obj (@adv @n @def @indef @pron @adj))
```

ルール番号1を含むリストにおいて、sは非終端記号であり、それに引き続くsub vpはsから導出される記号群である。その後の@のついた記号からなるリストは、先読み集合である。@の付いた文字列は品詞である。この集合は構文解析時に利用するが、この集合自身は、非終端記号から導出される記号群の先頭の記号の最左導出を繰り返すことにより得られるものである。構文解析部では、この集合により、どの構文解析ルールを選択するかを判断する。この例に現われた記号の意味は以下のようである。記号に続く":" の後にその意味を示した。

s : 文 sub : 主語 vp : 動詞 com : 補語
obj : 目的語 @n : 名詞 @def : 定冠詞 @indef : 不定冠詞
@pron : 代名詞 @adj : 形容詞 @adv : 副詞

現在このルールの数は少ないが、中学英語の基本5文型の文を解析できる程度のルールを用意している。用意した全てのルールは付録に示した。

4 語の文法機能の高速な判定のための不定知識の推定

4.1 語の形態知識の利用

構文解析を完全に行なうためには、文中の語の品詞をすべて決定する必要がある。語の品詞を決定するために辞書を検索するが、辞書の規模が大きくなると、語の検索に時間が費やされる。そのため、辞書の規模はできる限り小さくしたい。ところで語の形態の特徴により、語の品詞を推定できる場合がある。われわれはこれらの知識を利用して、辞書に登録する見出し語の数を減らしている。

以下に語の語尾の特徴により語の品詞が推定されるいくつかの品詞推定ルールの例を示す。

1. -sion,-tion であるとき、名詞とする
   例えば、conversion,production など
2. -ness,-ment であるとき、名詞とする
   例えば、happiness,treatment など
3. -tive,-ous,-able であるとき形容詞とする
   例えば、possitive,delicious,capable など
4. -ly であるとき、副詞あるいは形容詞にする
   例えば、beautifully,gently など
5. -go,-get であるとき、動詞とする
   例えば、undergo,forget など
6. -ize,-ise であるとき、動詞とする
   例えば、realize,systematise など
7. -'s であるとき、形容詞あるいは名詞にする
   例えば、Ken's,Kumi's など

このルール 7. は名詞の所有格を形容詞あるいは名詞にするということである。Let's, He's, She's, It's のような単語は、3.2節で述べたように、それぞれ特別な品詞として、辞書に登録してある。

たとえば、apportion のような単語は、語尾が tion であるが、その品詞は名詞ではなく動詞である。プロトタイプシステムでは、もし、このような単語が辞書に登録されていないと、誤った品詞認定をすることになる。このような単語に対する処理は、今後の重要な検討課題である。

4.2 文構造による品詞認定ルール

文構造による品詞認定ルールは、辞書引きや語の形態知識の利用により文を構成するそれぞれの単語の品詞認定処理が一通り行なわれた後に、品詞認定がなされていない単語に対して適用される。その語が文の先頭または文末の単語であるのかないのか、あるいは前後の単語の品詞を参考にして、未決定

の品詞を決定する。もし品詞が決定できなければ、対象としている英文を構文解析することができないことになる。

このルールは、「品詞未決定語の前の語の品詞あるいは記号"hd."」、「品詞未決定語」、「品詞未決定語の後の語の品詞あるいは記号"term."」をリスト形式にしたものに、品詞未決定語の品詞を加えてリスト形式にしたものである。"hd." は品詞未決定語の位置が文頭であること、"term." は文末であることを示す。

これらのルールの簡単な説明をしておく。@の付いた文字列は品詞を表す。

1. 品詞未決定語の前後の語の品詞が @def（定冠詞）と @n（名詞）である場合には、品詞未決定語の品詞は @adj（形容詞）とする。

2. 品詞未決定語の前後の語の品詞が @indef（不定冠詞）と @n（名詞）である場合には、品詞未決定語の品詞は @adj（形容詞）とする。

3. 品詞未決定語の前の語の品詞が @prep（前置詞）で、品詞未決定語の文中での位置が文末である場合には、品詞未決定語の品詞は @n（名詞）とする。

4. 品詞未決定語の前の語の品詞が @def（定冠詞）で、品詞未決定語の文中での位置が文末である場合には、品詞未決定語の品詞は @n（名詞）とする。

5. 品詞未決定語の前の語の品詞が @indef（不定冠詞）で、品詞未決定語の文中での位置が文末である場合には、品詞未決定語の品詞は @n（名詞）とする。

6. 品詞未決定語の前の語の品詞が @adj（形容詞）で、品詞未決定語の文中での位置が文末である場合には、品詞未決定語の品詞は @n（名詞）とする。

7. 品詞未決定語の前後の語の品詞が @def（定冠詞）と@prep（前置詞）である場合には、品詞未決定語の品詞は @n（名詞）とする。

8. 品詞未決定語の前後の語の品詞が @indef（不定冠詞）と@prep（前置詞）である場合には、品詞未決定語の品詞は @n（名詞）とする。

これらの 1.～ 8. のルールの内部表現形式を以下にまとめる。リスト中の文字列 not-known は品詞未決定語を意味する。現在のところ、記号"hd." の入ったルールはない。

```
1. ((@def not-known @n) @adj)
2. ((@indef not-known @n) @adj)
3. ((@prep not-known term.) @n)
4. ((@def not-known term.) @n)
5. ((@indef not-known term.) @n)
6. ((@adj not-known term.) @n))
7. ((@def not-known @prep) @n)
8. ((@indef not-known @prep) @n)
```

5 プロトタイプシステムの実行例

5.1 システムの使用法

作成したプロトタイプシステムは、検索対象ファイル、検索語と文法機能の3つの引数を与えると、語句検索を行なう。検索語を含む文を切り出して構文解析を行ない文法機能のチェックを行なう。チェックに通ればそれを出力する。

検索対象ファイルには、ニューホライズン中学英語教科書の学年毎のテキストファイルを用意している。すなわち、3学年分それぞれのテキストファイルがある。文法機能の指定は次の7種類の入力パターンのなかでどれか一つを選択する。

1. 検索語の品詞
   指定した品詞として使われている検索語を検索する。
2. ((prev 検索語の一つ前の語の品詞) 検索語の品詞)
   検索語の前の語の品詞を指定して検索する。
3. ((next 検索語の一つ後の語の品詞) 検索語の品詞)
   検索語の後の語の品詞を指定して検索する。
4. ((prev-next 検索語の一つ前の語の品詞 検索語の一つ後の語の品詞) 検索語の品詞)
   検索語の前と後の語の品詞を指定して検索する。
5. ((forward 検索語の前方の一つの語の品詞) 検索語の品詞)
   検索語の前方に指定の品詞の語があるものを検索する。
6. ((backward 検索語の後方の一つの語の品詞) 検索語の品詞)
   検索語の後方に指定の品詞の語があるものを検索する。
7. ((for-backward 検索語の前方の一つの語の品詞 検索語の後方の一つの語の品詞) 検索語の品詞)
   検索語の前方と後方に指定の品詞の語があるものを検索する。

検索語の前方とは、検索語の一つ前の語から文の先頭に位置する語までが範囲であり、検索語の後方とは、検索語の一つ後の語から文末の語までが範囲である。

入力パターン 1. 以外のパターンはすべてリスト形式である。入力パターン 2. から 7. までは入力パターンリストに特殊な文字列（位置記号）が含まれている。この位置記号とそれに続く品詞のリストにおいて、位置記号は、その品詞が検索語に対してどのような位置にある語の品詞であるかを示す。それぞれの位置記号を表2にまとめておく。

位置記号	意味
prev	検索語の一つ前の語の品詞
next	検索語の一つ後の語の品詞
prev−next	検索語の一つ前の語の品詞と検索語の一つ後の語の品詞
forward	検索語の前方の一つの語の品詞
backward	検索語の後方の一つの語の品詞
for−backward	検索語の前方の一つの語の品詞と検索語の後方の一つの語の品詞

表2．位置記号とその意味

次のものは機能の入力パターンの例である。@の付いた文字列は品詞である。

```
1.@v                              :検索語が動詞として使われている文の検索
2.((prev @indef) @n)              :検索語が名詞で、その直前に不定冠詞のある文の検索
3.((next @def) @v)                :検索語が動詞で、その直後に定冠詞のある文の検索
4.((prev-next @pron @def) @v)     :検索語が動詞で、その前後に代名詞と定冠詞のある文の検索
5.((forward @pron) @v)            :検索語が動詞で、その前方に代名詞のある文の検索
6.((backward @n) @v)              :検索語が動詞で、その後方に名詞のある文の検索
7.((for-backward @pron @n) @v)    :検索語が動詞で、その前方と後方に代名詞と名詞のある文の検索
```

5.2 システムの実行例

いくつかの実行例を示す。

[実行例1]

```
>(find-sentence)

FILE-NAME?--->h1-1.sentence
WORD?--->cousin
WHAT-PATTERN-OF-FUNCTION?--->@n

-------------SHE IS KUMI~S !COUSIN.-------------
((SHE @PRON) (IS @V) (KUMI~S @ADJ) (!COUSIN. @N))
GOOD
>>> SHE IS KUMI'S !COUSIN.
USED-RULE=((KUMI~S including ~, as adjective or noun))
FINISH
NIL

>
```

>はkclのプロンプトである。システムを起動させるために、find−sentenceという名前の関数を呼び出す。FILE−NAME?、WORD?、WHAT−PATTERN−OF−FUNCTION?の表示で、検索者は検索対象ファイル、検索語、文法機能の入力をする。この例では、入力がそれぞれ h1−1.sentence, cousin, @n となっている。h1−1.sentenceは中学1年生の教科書ファイルである。

次に、切り出された文と文中の単語にその品詞を付けたものをリスト形式にして、切り出した構文解析の対象となる文を表示する。文字"!"を付けた単語(!COUSIN)があるが、それは、対象の文中での検索語の位置を明示するためである。GOODとは、検索機能が @n（名詞） であるので、品詞が名詞である検索語cousinを含む英文が構文解析に成功したことを意味する。つまり、この英文を検索者に

とって望ましい英文であると判断したことになる。もし、検索語を名詞として構文解析に失敗すると、NOGOOD を表示する。

検索者にとって望ましい英文であると判断すると、すでに表示した品詞付きの英文のリスト形式ではなく、通常の英文を表示する。望ましくない英文であると判断すると、何も表示しない。

USED−RULE の部分は、どのような語の形態知識や文構造による品詞認定ルールが使われたのかを表示をする。品詞を認定するための知識やルールが使われていなければ、NIL が表示される。この例では、この USED−RULE から、Kumi's を形容詞あるいは名詞にしたのがわかる。Kumi's を形容詞として解析に成功しているので、名詞としての解析は行なわない。

FINISH で終了を示す。

[実行例 2]

```
>(find-sentence)

FILE-NAME?--->h1-3.sentence
WORD?--->gold
WHAT-PATTERN-OF-FUNCTION?--->@adj

-------------HE DREW !GOLD STARS ON A BLUE FIELD.-------------
((HE @PRON) (DREW @V) (!GOLD @ADJ) (STARS @N) (ON @PREP) (A @INDEF)
 (BLUE @ADJ) (FIELD. @N))
GOOD
>>> HE DREW !GOLD STARS ON A BLUE FIELD.
USED-RULE=((((@ADJ NOT-KNOWN TERM.) @N)))
FINISH
NIL

>
```

検索者は検索対象ファイルに h1−3.sentence 、検索語に gold 、文法機能として @adj（形容詞）を入力する。h1−3.sentence は中学 3 年生の教科書ファイルである。この実行例の USED−RULE から、辞書未登録のため品詞が認定されなかった文末の単語 field の品詞が、その直前の単語の品詞が形容詞であるため、名詞として認定されたことがわかる。

単語 on は前置詞以外に副詞の品詞も有するが、それは前置詞としての解析に失敗した場合に試みられる。

[実行例 3]

```
>(find-sentence)
```

```
FILE-NAME?--->H1-2.SENTENCE
WORD?--->studied
WHAT-PATTERN-OF-FUNCTION?--->((for-backward @pron @adv) @v)

-------------WE !STUDIED JAPANESE AT SCHOOL.-------------
((WE @PRON) (!STUDIED @V) (JAPANESE @ADJ) (AT @PREP) (SCHOOL. @N))
NOGOOD
USED-RULE=NIL
-------------WE !STUDIED JAPANESE AT SCHOOL.-------------
((WE @PRON) (!STUDIED @V) (JAPANESE @N) (AT @PREP) (SCHOOL. @N))
NOGOOD
USED-RULE=NIL
-------------I !STUDIED FRENCH THERE.-------------
((I @PRON) (!STUDIED @V) (FRENCH @N) (THERE. @ADV))
GOOD
>>> I !STUDIED FRENCH THERE.
USED-RULE=NIL
-------------MIKE AND PAUL !STUDIED FOR THREE HOURS.-------------
((MIKE @N) (AND @CONJ) (PAUL @N) (!STUDIED @V) (FOR @PREP) (THREE @N)
 (HOURS. @N))
NOGOOD
USED-RULE=NIL
-------------MIKE AND PAUL !STUDIED FOR THREE HOURS.-------------
((MIKE @N) (AND @CONJ) (PAUL @N) (!STUDIED @V) (FOR @PREP) (THREE @ADJ)
 (HOURS. @N))
NOGOOD
USED-RULE=NIL
FINISH
NIL

>
```

検索者は検索対象ファイルに h1−2.sentence 、検索語に studied 、文法機能として ((for-backward @pron @adv) @v) を入力する。h1−2.sentence は中学 2 年生の教科書ファイルである。入力された機能のパターンは、検索語の前方の一つの語の品詞、検索語の後方の一つの語の品詞、検索語の品詞である。入力された機能から、システムは、検索語の前方に品詞が @pron（代名詞）の単語、検索語の後方に品詞が @adv（副詞）の単語を含んでいる文を探し、検索語の品詞を @v（動詞）として、その文の構文

解析を行なう。この実行例では検索語を含んでいる文が3つあって、その中で、検索語の前方に品詞が代名詞の単語、検索語の後方に品詞が副詞の単語をもつ構文として構文解析に成功したのは1つだけであったことがわかる。

6 考察と課題

検索語とその文法機能をあわせて検索できるようになると、例えば、コーパスから動詞の have, take, give を含む文を抽出し、それらの動詞が命令形として存在する頻度の比較をしたり、あるいは動詞の後に名詞がくる文のみを抽出し、それぞれの動詞の特徴を、動詞の後にくる名詞により、分析するなどの試みが可能になると思われる。また、英文コーパスを辞書にある例文の集まりとして見ることにより、英作文での単語の用い方が適切であるのかどうかの確認もできる。

構文解析できる英文は現在のところ基本5文型から生成される文のみである。様々な英文の型を構文解析できるように、構文解析ルールの拡充が必要である。

品詞が認定されていない単語に用いられる語の形態知識や文構造による品詞の認定ルールがあるが、この知識は辞書の規模を抑え、辞書を引かない有効な品詞推定の方法として今後も大いに利用されると考えられる。ただし、語の形態知識の利用で、誤った品詞の推定をしないように適切な処理が行なわれなければならない。

ところで、複雑な構造の文を含むコーパスのような英文データベースを検索対象とするとき、検索語を含む文中のすべての単語の品詞を認定しなければ構文解析できない、というのは大変非効率的である。したがって、文中の単語の品詞がたとえすべて認定されていなくても、部分的な構文解析によって、検索者にとって望ましい英文かあるいはそうでないのかを判断できるような方法の研究が望まれる。

参考文献

[1] 舘 清隆: "英語例文検索システムの公開利用に向けて", 福井大学情報処理センター NETWORK Vol.1,No.4,pp.9-25(1988.1)

[2] 舘 清隆: "文と前提", 近代文藝社 (1993)

[3] R. セジウィック: "アルゴリズム", 近代科学社 (1992)

[4] 勝俣: "新英和活用大辞典", 研究社 (1958)

[5] 中島 文雄: "英語の構造、上・下", 岩波新書 (1980)

[6] Karin,A. and Bengt,A.: "English Corpus Linguistics",LONGMAN(1991)

[7] Roger,G. et al.: "The Computational Analysis of English (A Corpus-Based Approach)",LONGMAN(1987)

付録

品詞・統語機能の記号表現

辞書、構文解析ルールで用いる品詞、統語機能記号は、次のようである。記号に続く":" の後にその意味を示した。@の付いた文字列は品詞のことである。

s	: 文	sub	: 主語	vp	: 動詞	com	: 補語
obj	: 目的語	adv	: 副詞（句）	prep	: 前置詞	obj1	: 目的語
np	: 名詞（句）	com1	: 補語	np1	: 名詞（句）		
@n	: 名詞	@adj	: 形容詞	@adv	: 副詞	@v	: 動詞
@aux	: 助動詞	@pron	: 代名詞	@prep	: 前置詞	@conj	: 接続詞
@intj	: 間投詞	@def	: 定冠詞	@indef	: 不定冠詞	@intr	: 疑問詞
@past	: 過去分詞	@nv	: 主語＋動詞	@naux	: 主語＋助動詞	@rp	: 関係代名詞

過去分詞は辞書に含めたが現在分詞は辞書に含めていない。現在分詞である語の形態は語尾に ing を含むので、この特徴により語の品詞を現在分詞と認定することができるためである。本研究では、He's や I'm などの主語と動詞があわさったもの、I'd や She'll などの主語と助動詞があわさったものを、それぞれ省略された形の語の品詞としている。また、be 動詞や have の語を用いた He's や I've などの省略された形の語の品詞は、主語と動詞あるいは主語と助動詞があわさったものとして扱っている。

LL(1) 法で用いた構文解析ルール

構文解析ルールはリスト形式で表現する。リストの最初の要素の数字はルール番号、次の要素は非終端記号で、その後に続く要素はその非終端記号から導出される記号（群）である。最後の要素は先読み集合のリストである。

```
(1 s sub vp (@n @def @indef @pron @adj))
(2 s sub vp com (@n @def @indef @pron @adj))
(3 s sub vp obj (@adv @n @def @indef @pron @adj))
(4 s sub vp obj adv (@adv @n @def @indef @pron @adj))
(5 s sub vp obj prep (@adv @n @def @indef @pron @adj))
(6 s sub vp obj com (@n @def @indef @pron @adj))
(7 s sub vp obj obj1 (@n @def @indef @pron @adj))
(8 sub @pron (@pron))
(9 sub np (@n @def @indef @adj))
(10 com obj1 (@n @def @indef @adj))
```

```
(11 com com1 (@adj @adv))
(12 com1 @adj (@adj))
(13 com1 @adv com1 (@adv))
(14 obj @pron (@pron))
(15 obj obj1 (@n @def @indef @adj))
(16 obj1 np (@n @def @indef @adj))
(17 adv @adv (@adv))
(18 adv @adv adv (@adv))
(19 prep @prep @pron (@prep))
(20 prep @prep np (@prep))
(21 np np1 (@n @def @indef @adj))
(22 np np1 @prep np (@n @def @indef @adj))
(23 np np1 @conj np (@n @def @indef @adj))
(24 np np1 @prep np @conj np (@n @def @indef @adj))
(25 np1 @n (@n))
(26 np1 @n @n (@n))
(27 np1 @def @adj @n (@def))
(28 np1 @indef @adj @n (@indef))
(29 np1 @indef @adj @n @n (@indef))
(30 np1 @indef @adj @adj @n @n (@indef))
(31 np1 @def @n (@def))
(32 np1 @indef @n (@indef))
(33 np1 @indef @n @n (@indef))
(34 np1 @adj @n (@adj))
(35 np1 @adj @n @n (@adj))
(36 vp @v (@v))
(37 vp @adv @v (@adv))
(38 vp @v @prep np (@v))
(39 vp @v @prep @pron (@v))
```